NDIIPP Project Proposal: Collection and Preservation of At-Risk Digital Geospatial Data

Abstract

The proposed joint project of the North Carolina State University Libraries and the North Carolina Center for Geographic Information and Analysis will focus on collection and preservation of digital geospatial data resources from state and local government agencies in North Carolina. “Digital geospatial data” consists of digital information that identifies the geographic location and characteristics of natural or constructed features and boundaries on the earth. Such data resources include geographic information systems (GIS) data sets, digitized maps, remote sensing data resources such as digital aerial photography, and tabular data that are tied to specific locations. Geospatial data are created by a wide range of state and local agencies for use in applications such as tax assessment, transportation planning, hazard analysis, health planning, political redistricting, homeland security, and utilities management. State and local data resources are in general of greater detail and more current than data available from federal agencies. Since production points for these resources are so diffuse—92 of 100 North Carolina counties have GIS systems, as do many cities—they are generally not addressed by data archival efforts at the federal level.

Earlier this year, the North Carolina Geographic Information Coordinating Council established a vision and implementation plan to “organize the geographic information assets statewide” under a program called NC OneMap. NC OneMap complements efforts of *The National Map*, the Geospatial One Stop, and the interests of the Federal Geographic Data Committee. One of the stated goals of NC OneMap is that “historic and temporal data will be maintained and available.”

Although many of the targeted data resources are updated on a frequent basis—daily or weekly—data dissemination practices focus almost solely on providing access to current data. While snapshots of older versions of data may be stored in agency archives, access is almost as a rule not available and there is, in general, no commitment to long-term preservation of the data or to time series creation. These complex data objects do not suffer well from neglect; long-term preservation will involve migration of data to supported data formats, media refresh, and retention of critical documentation. Emerging data streaming technologies further threaten archive development as it becomes easier to get and use data without creating a local copy—the secondary archive having been in part a by-product of providing data access.

The objectives of the proposed project include:

- Identification of available resources through the NC OneMap data inventory;
- Acquisition of at risk geospatial data, including static data such as digital orthophotos as well time series data such as local land records and assessment data;
- Development of a digital repository architecture for geospatial data, using open source software tools such as DSpace;
- Enhancement of existing geospatial metadata with additional preservation metadata, using Metadata Encoding and Transmission Standard (METS) records as wrappers;
- Investigation of automated identification and capture of data resources using emerging OpenGIS Consortium specifications for client interaction with data on remote servers; and
- Development of a model for data archiving and time series development.

Although this project will focus solely on the state of North Carolina, it is expected to serve as a demonstration project for data archiving and time series development elsewhere. This proposal is a unique opportunity to advance the digital preservation interests of the geospatial community and the Library of Congress with the involvement of GIS practitioners in local, state, and federal government.

I. Work Plan

I.A Content Identification and Selection

This collaborative project of the North Carolina State University Libraries and the North Carolina Center for Geographic Information and Analysis (CGIA) will focus on digital geospatial data, which may be defined as information that identifies the geographic location and characteristics of natural or constructed features and boundaries on the earth.[1] Geospatial data resources include geographic information systems (GIS) data sets, digitized maps, remote sensing data resources such as digital aerial photography, and tabular data that are tied to specific locations. This data may be displayed within GIS software or, in the case of digital images, in digital image processing software. Geospatial data are created by a wide range of state and local agencies for use in applications such as tax assessment, transportation planning, hazard analysis, health planning, political redistricting, homeland security, and utilities management. Although often created with specific applications and functions in mind, these data resources are used in applications ranging far beyond those initially intended. End-user historical applications that might make use of historical and time series data include analyses of urbanization processes, environmental change, demographic change, and land use change. Libraries might also use these data for the purpose of "geo-enabling" library resources, enabling place-based discovery of information resources through mechanisms such as gazetteer lookup.

County and municipal data resources are in many ways analogous to the Sanborn Fire Insurance Maps published at the turn of the last century. Those maps, while created with a very narrow purpose in mind, survived by virtue of their relatively stable analog form and the intervention of interested organizations, including the Library of Congress.[2] The new local geospatial data, while initially created for very specific administrative and operational purposes, already find uses in a wide range of applications beyond the intended uses. It has been estimated that the cost to create the geospatial data that exists at present for North Carolina counties and cities would be roughly $162 million.[3] Retrospective creation of a similar data collection for 1999 or 2000, for example, would be impossible. In the absence of legal and technical guidelines and the resources for agencies to preserve this data, each year vast quantities of it are being lost forever.

Content Subject Areas, Display, and Use Characteristics

Major types of data resources to be acquired include:

Vector data – These data resources model features on the earth's surface as points, lines, or polygons. For example, a well location or a school may be modeled as a point; a stream or street centerline may be modeled as a line; and a land parcel or school district may be modeled as a polygon. A vector data set may form a "data layer," such as a streets dataset covering a county. Vector datasets are not digital maps—maps are just one possible output—rather these data may be displayed or analyzed in many different ways, together with other data layers or inputs, using GIS software. State agency data is

[1] "Executive Order 12906 -- Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure", 59 Federal Register, April 13, 1994. Available (GPO Access) http://www.gpoaccess.gov/fr/index.html.

[2] "Sanborn Fire Insurance Maps." Available http://www.lib.berkeley.edu/EART/snb-intr.html.

[3] "Funding Allocation for State Data Production: A Cost Allocation Model," National States Geographic Information Council. Available http://www.fgdc.gov/I-Team/library/financial.html.

typically of larger scale (more detailed) than federal or national-level data. Even more detailed vector data is available from the county and municipal government level. The wide range of data layers available from county and municipal governments is illustrated by a sampling of typical layers shown in the table below. (See Appendix A for a detailed description of vector data, with examples.)

Table 1. Data layers often available from county or municipal agencies

Physical Environment Geodetic Land Cover Soils Elevation Watersheds Surface Waters **Land Use** Zoning Districts Buildings Structures Parcels, including Public Land Use	**Infrastructure** Water Distribution Lines Roads Sewer Lines Railroads Schools Airports Hospitals Cell Towers Prisons Landfills **Remote Sensing** Aerial Imagery	**Emergency and Disaster Response** Fire Districts Emergency Mgmt. Districts Flood Zones **Political and Jurisdictional Boundaries** Voting Districts County Lines Municipal Lines School Districts Extra-territorial Jurisdictions

Digital orthophotos – A conventional aerial photograph contains image displacements caused by camera lens distortion, camera tip and tilt, terrain relief, and scale. The effects of camera tilt and terrain relief may be removed through a rectification process to create a computer file referred to as a digital orthophoto, which is a uniform scale photographic image—essentially a photographic map. Through a process known as "heads-up digitizing," a data user may edit or create vector data layers on top of digital orthophotos, which display features that may be omitted or generalized on other cartographic maps.

County government-produced orthophotos are typically created at resolutions ranging from six inches ground surface per pixel in urban areas to two feet per pixel in rural areas. In North Carolina, 80 counties currently have digital orthophotos, and 5 additional counties will have these by 2006. Counties typically conduct orthophoto flights every two to five years. File sizes for an individual county flight can total as much as 100 gigabytes or more, and 40 counties have had multiple flights.[4] Overall frequency of orthophoto flights is increasing as is the quantity of data generated in each subsequent flight. Statewide orthophotos at a lower resolution (one meter) have also been created through a combined state and federal effort for the years 1993 (black and white) and 1998 (color infrared), with actual image dates spanning several years. (See Appendix A for a detailed description of digital orthophotos.)

Digital Maps – Less common are digital maps which may be georeferenced in the same way as digital orthophotos. Georeferenced digital maps are typically created at the state agency level, and include digital raster graphics (digital topographic maps) and travel maps. Additionally, a range of non-georeferenced digital maps are available from state and local agencies, including traffic volume maps, bridge maps, county road maps, bike route maps, and zoning maps. Non-georeferenced digital maps will be acquired selectively, taking into account the potential for such maps to be used in geoprocessing applications. (See Appendix A for a detailed description of digital maps, with examples.)

[4] Flight data derived from combined North Carolina Department of Transportation, North Carolina Department of Agriculture, and NCSU Libraries inventories. 2005-2006 flight plans for many counties remain unknown.

Tabular Data – Tabular data are numeric or textual data stored in database, spreadsheet, comma separated value, or other like formats. Some tabular data may be associated with geographic features such as land parcels or census tracts. This project will focus solely on non-federal tabular data occurring at a granularity greater than county level. This largely concerns tax assessment data, which associates land parcel records with such attributes as property value, purchase price, purchase date, building type, construction date, square footage, zoning, land use, and owner name. These data are produced by county tax assessment agencies and are made publicly available according to public records law, often bundled together with land parcel vector data. (See Appendix A for a detailed description of tabular data, with examples.)

Processes, Systems, and Tools Used to Identify and Select Content

A key impediment to ongoing acquisition of state and local data has been the absence of a current inventory of the rapidly evolving body of data resources available from state and, especially, local agencies. Since the mid-1990s, a number of efforts have been undertaken to track the spread of GIS activity among local government agencies. Such efforts have typically been incomplete given the complexity of the task of surveying 100 counties and 140 cities. Furthermore, such inventories become obsolete quickly.

This project will benefit from a key and timely initiative put forward by the North Carolina Geographic Information Coordinating Council (GICC), which in November 2002 adopted a recommendation to develop an on-going inventory of municipal, county, state, and federal data. CGIA was designated to take the lead on the effort, and the NC OneMap Data Inventory was initiated in September 2003. This survey will act as a living inventory, as distinct from earlier one-time and quickly outdated static inventories. The initial Fall 2003 survey is making use of the SurveyMonkey commercial Internet service. In 2004, in connection with the proposed project, CGIA will develop a next-generation survey software system, using open source software tools, and will make the inventory information available to the NCSU Libraries on an ongoing, dynamic basis for use in data identification, selection, and acquisition efforts.

Additional survey resources that will be available for identification of content include:

North Carolina Flood Mapping Project – This project, initiated in September 2000, was kicked off with an investigation into data availability of local government data needed to update the state's flood hazard maps in the aftermath of Hurricane Floyd. The survey process, which was conducted by CGIA, included site visits to relevant local agencies. The NCSU Libraries and North Carolina Department of Transportation provided input from earlier, separate investigations into data availability.

NCSU Libraries Local Government Data Acquisition Project – This NCSU project, initiated in spring 2000, resulted in an inventory of data holdings for more than one-half of the state and in acquisition of data for 47 counties in North Carolina. (See Appendix B.) The NCSU Libraries continues to maintain a comprehensive list of links to web sites, map servers, and data download sites at county and municipal agencies.

North Carolina Dept. of Transportation (NCDOT) – Since the late 1990s, the NCDOT has worked to acquire orthophotos and vector data sets such as street data from most North Carolina counties. NCDOT and NCSU have collaborated on local government data and inventory sharing since 2000.

North Carolina Dept. of Agriculture – Since 2001 the Department of Agriculture has worked to acquire land parcel data from a majority of counties in the state. The Department of Agriculture and NCSU Libraries currently share information on data inventories and holdings.

1997 National Geospatial Data Framework Survey – More than two hundred counties, municipalities, state government agencies and federal agencies responded in the North Carolina component of this survey, administered by CGIA. Although now largely historical in nature, this survey provided the basis for the NCSU Libraries data acquisition work begun in 2000 and still provides useful information about older local data holdings.[5]

Additional content will be obtained through systematic analysis of appropriate county, municipal, state, and lead regional organization web sites and by direct contact with relevant agencies as appropriate.

An expected benefit of this project will be the accumulation of inventory information from several different sources, creating the opportunity to provide feedback to the NC OneMap Data Inventory effort with regard to completeness of information acquired through the new online survey process. Data availability and acquisition progress will be tracked in a GIS database, allowing for query and display of inventory information, acquisition status, and data holdings by county, municipality or other place within the state.

Legal and Intellectual Property Issues

State and local government data resources in North Carolina are subject to public records law and as such must be made available to the public "free or at minimal cost unless otherwise specifically provided by law," with a further stipulation that "minimal cost is defined as 'the actual cost of reproducing the public record or public information.'" According to NC Gen. Statutes Section 132-10, "Qualified exception for geographical information systems," agencies reserve the right to restrict resale of the data and to restrict commercial use, with some exceptions for the real estate industry.[6] In practice, freedom of access to data varies greatly, with some counties providing free public download and others charging considerable sums for data access. One of the concerns agencies have about access to older versions of data is the issue of liability with regard to possible confusion of current and old data. Many counties require individuals accepting data to sign a liability waiver. Municipal and state agencies are generally less restrictive in terms of data access, though some state agency data is restricted for privacy (e.g., public health and livestock disease data) or preservation (e.g., endangered species or cultural heritage) reasons.
In the short term, county government restrictions may require select county data resources to be restricted from redistribution or, in fewer cases, excluded from acquisition, but the general experience of data acquisition and partnering over the past three years has been that, as counties become more comfortable with data distribution, relevant policies become more relaxed. The growing openness towards making data freely available is reflected in the increasing number of counties providing free public download and the growing number of data sharing agreements being obtained in connection with the NC Floodplain Mapping Project. (See Appendix B.) Legal issues to be addressed in retention include downstream adherence to government restrictions on data redistribution and addressing the issue of disclaimer presentation as a precursor to data access.

[5] "NSGIC/FGDC Framework Survey: North Carolina." Available http://fgdc.er.usgs.gov/framework/survey_results/samples/html/northcarolina.html.

[6] "North Carolina Public Records Law." Available http://www.ncpress.com/publicrecordslaw.html.

Issues Affecting Technical Sustainability Over Time

Data formats – Vector data are typically stored in commercial data formats such as the Environmental Systems Research Institute (ESRI) "shapefile" and "coverage" formats, while new formats such as the ESRI "geodatabase" are emerging. Non-proprietary exchange file formats such as the Spatial Data Transfer Standard (SDTS) have not taken root in the industry, though open standards for client access to data servers are gaining ground. While ESRI formats predominate in North Carolina, data are available from some counties in formats such as the Intergraph or Understanding Systems OASIS formats. In the proposed project, data in non-ESRI formats will be converted to ESRI formats, with both the original and derived file retained. At some future point, these data will need to be migrated to newer formats as existing formats cease to be supported by successive software versions. Image products are typically produced as uncompressed TIFF or BIL files. Derivative MrSID and JPEG files are typically made for ease of distribution due to the large file size of the uncompressed imagery. Images will be acquired, whenever possible, as uncompressed files. Where no uncompressed file exists, images will be converted to TIFF format for archival purposes, with both the original and derived file retained.

Metadata and documentation – In order to use geospatial data reliably, the typical user will require metadata or technical documentation, which informs the user about data structure, content, georeferencing system used, data lineage (or processing history) and recommended use. Additional ancillary documentation such as data dictionaries for attributes (e.g., land use codes for land use polygons) may be required. The Federal Geographic Data Committee (FGDC) published the Content Standards for Digital Geospatial Metadata (CSDGM) in 1994, and federal agencies were mandated to begin using the standard in 1995.[7] The standard, which reached version 2 in 1998, has since been widely adopted the state government level, with a lower level of adoption at the local level. The state of North Carolina was an early adopter of the standard and CGIA has actively promoted the standard at the state and local level through grant-funded workshops and outreach.

Emerging data access technologies – It is often the case that secondary archives—such as those found in libraries--survive over time more readily than do those of the information producers. In the case of geospatial data, the secondary archive is often a by-product of the need to provide access. Over the past two to three years, new data delivery technologies allowing users to stream in data to their applications directly from the Internet without downloading data sets have begun to emerge. Available methods include ESRI's ArcIMS image and feature services, OpenGIS Consortium (OGC) Web Map Service (WMS) and Web Feature Service (WFS), and access to commercial data services via SOAP (Simple Object Access Protocol) connections. Such access methods are attractive to individuals and organizations, particularly because these data tend to be large in size and the user may only need to interact with a small subset. As such access methods become more widely used, the development of secondary archives may diminish, putting these resources at risk. On the other hand, these new technologies provide an opportunity to automate the processes of remote data inventory and acquisition and make the process of data archive development more sustainable.

The Content Selection and Identification phase will largely occur in year one of the project, but some components will continue through years two and three. Detailed content identification and selection plans for the three-year project may be found in Appendix A.

[7] Information on geospatial metadata standards is available on the Federal Geographic Data Committee Metadata website. Available http://www.fgdc.gov/metadata/metadata.html.

I.B Content Acquisition

Order of acquisition will be refined during the Content Identification and Selection phase, but some general principles are expected to apply as a default framework for acquisition efforts. These principles include:

- "time-sensitive" data—those used to create time series or at short-term risk of disappearing—will be acquired first. Targeted resources would include vector data that are routinely updated;
- digital orthophotos, while "at risk," are not "time-sensitive"; acquisition will be largely deferred to years two and three of the project. Since orthophotos comprise the bulk of total collection file size, this also allows time for the redundant storage model to be fully implemented and tested;
- county and state data will be targeted from the first year of the project, since there is greater knowledge about data availability for these agencies and more comprehensive benefit from initiation of time series development. Municipal, lead regional organization, not-for-profit, and university data will be targeted in later phases;
- geographic gaps in existing agency acquisition efforts will be targeted for early acquisition, with a particular focus on Western North Carolina, which has not been thoroughly addressed in earlier acquisition efforts.

Technical Specifications and Standards For Capture Mechanisms

Because of the variety of distribution methods provided by local and state agencies, a number of different approaches will be used to acquire data:

- direct download from data servers – 14 counties, 7 cities, and several state agencies in North Carolina currently allow direct FTP or web download from their servers. The number of such services has increased steadily each year. In order to develop time series in a reliable fashion agent tools for automated download will be explored;
- upload by data producers to NCSU Libraries servers – NCSU Libraries servers will be equipped to allow incoming file transfer of data by state and local agencies;
- external hard drive transfer of data – Currently preferred by some partners, this method is expected to become more common as agencies include external drive support in their technical infrastructure. Onsite visits for external drive transfer from CD-ROM will be conducted on an as-needed basis;
- web extract from agency interfaces – Several local agencies have developed elaborate Web interfaces to allow for download of individual data files for specific locations, while not enabling bulk download. Agent applications may be constructed for automated parsing of individual web sites, although this approach is not sustainable on a wide scale due to the site-specific nature of the tools developed;
- feature server extraction – 58 counties, 14 cities, and 4 state agencies in North Carolina currently maintain web map servers, which allow users to construct maps on the fly using ordinary Web browsers, interacting remotely with data on the agency server. The underlying data will be extracted from these servers where possible (see Appendix B);
- non-Magnetic media transfer – Data will also be transferred by CD and DVD where necessary. Orthophoto holdings for individual counties may require between 10 and several hundred CDs for full data transfer.

Content Authentication Issues

Vector and tabular data will be acquired directly from producing agencies. In an effort to extend time series into the past, vector and associated tabular data will also be acquired, where possible, from intermediaries who have in the past acquired the data (CGIA, NCDOT, and NC Department of Agriculture), and appropriate provenance metadata will be created.

Due to the size of uncompressed orthophoto collections and the cost of transfer, uncompressed orthophotos will be acquired, where possible, from intermediaries such as CGIA, NCDOT, and NC Department of Agriculture. Otherwise, uncompressed orthophotos will be acquired directly from producing county agencies.

Metadata Processing

FGDC Metadata – FGDC metadata is available for data originating from several state agencies. Availability of FGDC metadata for counties and cities is less comprehensive. Where FGDC metadata is available, it will be acquired along with the data. Where such metadata is not available, minimal data documentation will be gleaned from other agency documentation, through the NC OneMap Data Inventory, or by posing a set of questions to producers to acquire a minimal set of metadata. Batch extraction of selected metadata elements will also be undertaken. (For a discussion of tools see Appendix B.)

METS Metadata – While the FGDC metadata standard is exhaustive with regard to describing data content, structure, lineage, and georeferencing, it does not address some critical preservation metadata requirements. The proposed project will adapt the Metadata Encoding and Transmission Standard (METS)[8] to geospatial data for the following purposes:

- to create a wrapper for the various components that make up a complex data object, including: the one or more data files making up a data object, georeferencing files, metadata files, and ancillary documentation;
- to provide a container for administrative metadata that is external to the FGDC metadata. This includes: data acquisition history, acquisition process, rights pertaining to the archive (as distinct from rights as expressed by the data producer);
- to link to services that operate on the data object using the METS "behavior" component. This will provide an end user or harvester with a pointer to a web map or feature service operating on a different copy of the same data; and
- to provide Submission Information Package (SIP), Archival Information Package (AIP), and Dissemination Information Package (DIP) functions in a digital repository context.

Commercial and open source authoring tools for METS records will be evaluated, though it is anticipated that the project will need to assemble METS records through batch scripting and XSLT processes in the initial stages. Creation of a GIS profile for METS will be explored.

The application of METS to geospatial data would continue a tradition of technological cross-fertilization between the GIS and library communities. GIS industry adoptions from the library community include study of the library cataloging standards in connection with creation of the federal Content Standard for Digital Geospatial Metadata in 1994. A second example of technology crossover was the adoption of the Z39.50 standard for cross-database searching. Though the standard was developed within the library

[8] Metadata Encoding & Transmission Standard website. Available http://www.loc.gov/standards/mets/.

community, the protocol was adopted by the GIS community, which in 1995 created a new "geo" Z39.50 profile for use in development of the National Spatial Data Clearinghouse search system.[9]

The library community also stands to learn much from the GIS community. Meta-search systems were developed on a wide scale in the national and global GIS communities before becoming common in the library community. Extensive development of non-MARC FGDC metadata by the mid-1990s spawned a small industry of authoring tools, and led toward XML metadata database development work. The GIS community has also been wrestling with Web Services-based information delivery issues at early stages in development of Web Services technologies. In the context of digital preservation, the library community may also stand to learn from the current experience of the GIS community in management of large quantities of digital content and associated metadata at the federal government level.

Detailed acquisition plans for the three-year project may be found in Appendix B.

I.C Partnership Building

The proposed project will build upon an existing statewide organizational framework, key components of which are the North Carolina Geographic Information Coordinating Council (GICC), the North Carolina Center for Geographic Information & Analysis (CGIA), and the NC OneMap Initiative.

The GICC is established by legislation and is charged with improving the quality, access, cost-effectiveness and utility of North Carolina's geographic information and promoting geographic information as a strategic resource for the state.[10] The Council creates policy and resolves technical issues related to North Carolina geographic information and GIS systems and fosters cooperation among government agencies, universities, and the private sector.

The GICC includes 33 members representing the GIS community statewide. Ten members are from local government, and the academic community is represented by the President of the University of North Carolina system and the President of the NC Community College System. The legislation also established six committees that support the GICC, including an active Local Government Committee.

The GICC established one of the first cooperating partnerships with the Federal Geographic Data Committee (FGDC). Through this partnership, the GICC promotes the goals of the National Spatial Data Infrastructure and actively cooperates with US Geological Survey (USGS) in the development of the National Map and the Geospatial One-Stop program.[11] This formal coordination structure is helping North Carolina realize the full potential of a coordinated approach for the use of geographic information technology within the state.

The CGIA is the primary state GIS agency and serves as staff to the GICC. In this role, CGIA is responsible for implementing the goals and strategies of the GICC. Established in 1977, CGIA also operates a GIS service program and provides GIS services—application development, data development, spatial analysis, system planning, image analysis, and general GIS technical assistance—to users in North Carolina.

[9] Doug Nebert, "Use of Z39.50 to search and retrieve geospatial data." Available http://www.fgdc.gov/publications/documents/clearinghouse/dlipaper395.html.

[10] NC Geographic Information Coordinating Council website. Available http://cgia.cgia.state.nc.us/gicc/.

[11] National Map website. Available http://nationalmap.usgs.gov/.

For the next two years, the GICC has adopted as its first priority the design and implementation of a comprehensive statewide geographic data resource, called NC OneMap. NC OneMap is the North Carolina version of the National Map and will serve the basic information requirements for decision-making in the community, statewide, and in support of national priorities. Users will be able to view geographic data seamlessly across North Carolina; search for and download data for use on their own GIS; view and query FGDC compliant metadata; and determine who has what data through an online data inventory. A NC OneMap demonstration site (www.nconemap.org) is already operational. The demonstration site was built on the architecture of the National Map and invokes OGC specifications.

More than a dozen county and municipal governments are participating in the NC OneMap demonstration viewer, which combines geospatial data in real time from federal, state, and local government servers in a seamless map display. In addition, more than 40 local governments have signed data sharing agreements with CGIA. NC OneMap is the portal that links geospatial data from users across the state. The goal of the GICC is to achieve participation in NC OneMap by all 100 NC counties and more than 140 municipalities in addition to state and federal government GIS users in North Carolina. NC OneMap will serve as a model for USGS and FGDC efforts to promote the use of geospatial data nationally.

The GICC has formally adopted an implementation plan that includes:

- the adoption of data content standards built on national standards under the Geospatial One-Stop program;
- formal community data sharing agreements;
- a system design that addresses hardware requirements, system architecture, network protocols, security issues, and privacy concerns; and
- an inventory—already underway—of GIS users across the state that will document data holdings and data gaps.

This proposal from NCSU and CGIA to the Library of Congress presents an important opportunity to address the issue of digital geospatial data preservation in the early stages of NC OneMap as part of the implementation planning process. North Carolina is an ideal test ground for the following reasons:

- an active and robust coordination structure—the GICC and its committees—is in place;
- successful cooperation between state, federal, and local governments and the academic community is already a reality as formal agreements and partnerships for sharing geospatial are already in place;
- a resource for accessing and sharing geospatial data – NC OneMap – is already under development and a demonstration site is operational. The infrastructure and institutional framework of NC OneMap will also serve as the focal point and core mechanism for preserving digital geospatial data.

I.D Content Retention and Transfer

Content will be physically acquired by NCSU Libraries from data producing agencies and intermediary agencies and stored in an online, mirrored digital repository. It is expected that individual state and local agencies will concurrently expand roles with respect to development of long-term archives and time series.

Repository Architecture

A 12.6 terabyte near-line storage disk array will be implemented with RAID 5 fault-tolerant architecture. The equipment will be connected to existing NCSU Libraries gigabit Ethernet switches to support rapid disk writes over the network. A tape library solution will be implemented to provide backup and redundant offsite data storage. In addition, a redundant disk-based archival copy will be maintained at the NCSU Libraries' secure offside storage node, which is supported by library and university technical staff.

Tentative plans call for deployment of the DSpace digital repository software developed by Massachusetts Institute of Technology (MIT) and Hewlett Packard.[12] Although DSpace does not currently support METS records, version 1.2 is expected to support METS for Submission Information Package (SIP), Archival Information Package (AIP), and Dissemination Information Package (DIP) functions, which will be a requirement for use of DSpace in this project.[13] Analysis of emerging METS authoring tools will focus on both individual and batch authoring of METS records. Other digital repository software tools such as FEDORA (Federated Digital Object Repository Architecture), developed by the University of Virginia and Cornell, will be evaluated over the period of the project.[14] The proposed project work will leverage existing NCSU Libraries investments in infrastructure and expertise for digital repository development.

Research and Development

This project will involve an investigation into current data storage and preservation technologies and methodologies as currently employed by large federal repositories. Particular focus will be given to the EROS Data Center, which has demonstrated experience in managing extremely large geospatial data archives.[15] While the EROS Data Center is not believed to be currently addressing state and local data—focusing instead on national and global data—it is expected that important lessons may be learned from the EROS Data Center experience.

Post-Project Retention and Transfer of Content

The NCSU Libraries will commit to storing and preserving data acquired during the three-year period of the project (data translations) for a minimum of five years beyond the conclusion of the funding period, and will pursue additional sources of funding to support continuation beyond this period. Following the completion of the three-year project, it is expected that the project components and content areas will fall into one or more of the following categories:

a) Components that will have become part of state and local government data infrastructure and processes. The "Characteristics of NC OneMap" as defined by the North Carolina Geographic Information Coordinating Council include the statements that "Historic and temporal data will be maintained and preserved" and that "NC OneMap data are reliably maintained by the data provider organization through partners and formal arrangements."[16] During the period in which

[12] DSpace Federation website. Available http://www.dspace.org/.

[13] Announcement on dspace-tech listserv, Sept. 5, 2003. Available: http://mailman.mit.edu/pipermail/dspace-general/2003-September/000006.html

[14] Federated Digital Object Repository (FEDORA). Available: http://www.fedora.info/. Of particular interest with regard to FEDORA is the Web Services-based architecture and the potential to use the "behavior" component of METS records to link data objects to disseminators, or services, that operate on the objects in question (e.g., map renderers).

[15] EROS Data Center website. Available: http://edc.usgs.gov/.

[16] "NC OneMap Vision and Characteristics." Available: http://www.cgia.state.nc.us/nconemap/documents/visiondoc.pdf

the supporting infrastructure for NC OneMap objectives is put into place, a three-year span of time series data will have been acquired by the project. In addition to accumulated content, the proposed project will have developed a body of experience and serve as a test-bed for development of best practices with regard to long-term access and preservation. The involvement of CGIA, as the state's coordinating GIS agency, will serve to further disseminate best practices. North Carolina's history of leading-edge activities with regard to geospatial data infrastructure and close involvement in activities such as the National Map enable the project partners to leverage experience and knowledge of current practices onto the national scene.

b) Components that NCSU Libraries continues because of fit with organizational mission. The NCSU Libraries is committed to providing a high level of service with respect to geospatial data access, and there is a high level of university-wide interest in GIS, with over 35 academic departments involved in GIS activities. Local government data acquisition efforts, which have been carried out on a more limited scale to date, have helped to change the nature of GIS work done on campus, as the large-scale local data enables application development and pursuit of research in fields not previously using GIS. The NCSU Libraries will have a clear interest in continuing to provide long-term access to such data, including time series resources developed as a result of long-term archival processes.

c) Components which prove to be unsustainable. It is entirely possible that some of the project components will turn out to be unsustainable over a longer period of time. Following analysis of findings in an end-of-project evaluation, additional funding might be pursued to continue development in acquisition areas that show promise but which do not fall within the set of sustainable activities that the project partners and collaborators identify as part of their organizational scope. If any acquired content has not, in the five years following the end of project funding, been assimilated into ongoing NC OneMap and NCSU Libraries long-term archival functions, the content in question will be made available to the Library of Congress.

Detailed content retention and transfer plans for the three-year project may be found in Appendix D.

II. Staffing and Institutional Capacity

The proposed project benefits from being a collaborative effort among geospatial data repository managers at NCSU Libraries and the North Carolina Center for Geographic Information and Analysis. North Carolina State University is a research-extensive institution with well established strengths in science and technology as well as in statewide extension and local government partnerships. With its resident expertise in information science and commitment to long-term management and protection of resources, the NCSU Libraries is uniquely prepared to lead this project. With its experience in developing a statewide geospatial data archive, promoting standards and best practices for data production and documentation in the state's geospatial data user community, and in developing innovative tools for data access, CGIA will play an important role in identifying and acquiring content as well as in promoting development of historical archives and time series within the data producer community and relevant organizational entities.

II.A NCSU Libraries Technical and Organizational Capacity

Technical Architecture

The NCSU Libraries houses and supports a server, backup and network infrastructure that is comparable in sophistication and staffing expertise to many university computing centers. The library supports a high-availability failover server architecture in which there are no single points of failure for production services. Up to six terabytes of data are currently backed up through a tape library. The library also utilizes offsite storage for tapes and backup servers. The Libraries has a direct gigabit connection to the campus backbone. NCSU, a member of Internet 2, is also a founding member of the North Carolina Networking Initiative (NCNI). NCNI provides gigapop network capacity among the Triangle area universities and research community with links to the statewide North Carolina Research and Education Network. Among the digital library initiatives with which the NCSU Libraries is currently engaged is an NSF-funded collaborative project to build a web-based searchable archive of wood anatomy images (the majority of which will exist only in digital form) and associated metadata, with managed content expected to total 5-7 terabytes.

Geospatial Data Infrastructure and Expertise

The NCSU Libraries also brings considerable expertise with geospatial data resources to bear on the project. Since 1993 NCSU has maintained a data server for campus-wide access to data resources, and in 1997 the Libraries began to serve data through interactive web-based mapping applications. As a result of cumulative efforts in data collection building, access tools development, outreach, and partnering, in 2001 the NCSU Libraries became the first library to receive the industry ESRI Special Achievement in GIS Award. For the past decade, library staff have developed an ongoing collaboration with the Center for Earth Observation, which is based in NCSU's College of Natural Resources and manages numerous teaching and research programs. That center offers substantial expertise in geospatial technologies and has worked and consulted with the library on a number of grant-funded initiatives focused on public access to GIS data.

The NCSU Libraries has well developed expertise in the area of identifying, acquiring, and managing state and local agency data resources. Beginning in 2000, the library led a University Extension-funded effort to acquire county and municipal data in central and eastern North Carolina. This effort spurred development of expertise in the areas of data acquisition negotiations, metadata authoring, image compression, development of tools for batch processing of data and metadata, and large-scale data collection management. (See Appendix B for a summary of data acquisition efforts.)

II.B NCSU Libraries Project Staff

Primary Staff (existing)

Steven P. Morris (.15 FTE), Head, Digital Library Initiatives, will serve as project director and will be the authority on data archival issues. He will act as the principal investigator for NCSU Libraries on the project. In addition to coordinating overall project work, he will be responsible for directing development of digital repository infrastructure and will lead investigations in the area of OGC protocol-based data identification and extraction. Mr. Morris has led data services efforts at NCSU since 1997 and initiated and led an ongoing effort to acquire state and local data for archival purposes. He represents the University of North Carolina System on the State Mapping Advisory Committee (SMAC), a subcommittee supporting the GICC.

Jefferson F. Essic (.25 FTE), Data Services Librarian, Research and Information Services, will provide GIS technical expertise with regard to identification, evaluation, and processing of data resources. He will also integrate current library data inventory information and holdings into the project effort. Mr. Essic currently manages data services operations at NCSU. Previously, as an employee of Triangle J Council of Governments, Mr. Essic led a regional effort in the area of local government data acquisition and acted as Lead Regional Organization representative to the SMAC.

James M. Jackson Sanborn (.1 FTE), Metadata Architect/Assistant Head, Digital Library Initiatives , will serve as consultant on metadata architecture issues, including design of metadata schemas, development of tools, such as XSLT style-sheets for metadata transformations, and database design. Mr. Jackson Sanborn currently designs and develops XML and database applications related to information discovery and consults on metadata architecture issues in a variety of information resource projects within the library.

Jacqueline Samples (.1 FTE), Metadata Librarian, Cataloging Department, will supervise metadata production work with regard to creation of technical, administrative, and descriptive metadata needed for preservation purposes. Ms. Samples currently supervises non-MARC metadata production for the NCSU Libraries Special Collections content.

Secondary Staff (existing)

Robert W. Main III (.05 FTE), Technical Operations Manager, Systems Department, will design and implement the disk repository and backup architectures and incorporate them into the existing systems architecture of the library. He will supervise staff who maintain systems administration and daily operation functions to assure data integrity, security, and reliability of systems.

Barry Gaskins (.1 FTE), Applications Analyst Programmer II, Digital Library Initiatives Department, will do programming and development work on enhancing DSpace, as well as other repository software and tools that are implemented as part of the project.

Scott Devine (.05 FTE), Head, Preservation Department, will provide consulting and expertise on broader preservation issues and will facilitate transfer of expertise from other digital preservation initiatives in which the library is involved.

Project Funded Staff

Project Manager (1 FTE) - The position of Project Manager will be created following receipt of the grant. This person will take lead responsibility for identifying content for acquisition, leading data collection work, supervising FGDC and METS metadata production work, and managing flow of content into the repository.

Temporary Data Management Staff - North Carolina State University has numerous graduate students with the appropriate background and skills. The project will employ them to help coordinate with local agencies on acquisition of data files and metadata, acquire necessary documentation, carry out metadata file creation and metadata transformations, and conduct transfer of data and metadata into the repository.

II.C CGIA Technical and Organizational Capacity

CGIA is the state agency responsible for providing geographic information and services statewide, developing an Internet mapping and information system—NC OneMap—and serving as staff to the NC Geographic Information Coordinating Council. Operating since 1977, CGIA has completed more than 400 hundred projects with state agencies, counties, municipalities, nonprofit organizations, universities and businesses as well as managed some 100 GIS data layers for statewide use. CGIA staff has extensive experience in database management, data creation, custom computer applications, analysis, reporting, custom mapping, partnership building and professional presentations. CGIA analysts and programmers apply GIS tools to North Carolina data on a daily basis.

CGIA, a receipt-supported agency, has provided technical support to hazard mitigation planning and emergency management, built GIS capacity in public health agencies, created an inventory of conservation and open space, created tools for coastal land use planning, designed Internet mapping for rural telecommunications, and assisted the state's Floodplain Mapping Program. In addition to its expertise in GIS, CGIA's professional staff has training and experience in water quality, city and regional planning, economics, demographics, forestry, geography, geology, transportation, database management, information technology and system design.

Several of CGIA's recent projects have featured application development to customize desktop GIS and to create new GIS-based decision support tools. For the NC General Assembly, CGIA developed the "District Builder" system that enables users to create new voting districts statewide. CGIA is a partner with the NC Floodplain Mapping Program, the National Weather Service, and USGS in development of a flood inundation and forecast mapping system for the Internet. CGIA has also developed custom GIS projects for joint land use planning in the Fort Bragg region, and land suitability analysis for coastal counties. Equally important have been projects funded and assisted by federal partners relating to National Spatial Data Infrastructure, metadata, clearinghouse, the National Map, statewide aerial imagery and other efforts to build statewide GIS capacity.

II.D CGIA Project Staff

Primary Staff (matched salary)

Zsolt Nagy (288 hours/year), Program Manager, Geographic Information Coordination Program, will be the principal investigator for CGIA on the preservation project and provide oversight for NC OneMap partnering, inventory and infrastructure development. Mr. Nagy, who has worked with CGIA since 1982, has done considerable work on national, state, regional and local geographic information initiatives, including efforts to develop the National Spatial Data Infrastructure. He provides executive staff support to the North Carolina GICC and is integral to the NC OneMap initiative. Mr. Nagy is currently President-Elect of the National States Geographic Information Council (NSGIC).

Tom Tribble (96 hours/year), Manager, Asheville Field Office, will provide outreach to local agencies, playing a key role in enabling access to local government data resources. Mr. Tribble acts as lead CGIA liaison to the Local Government Committee of the North Carolina GICC. He has played a key role in building CGIA's capabilities and reputation since the mid-1980s.

David Giordano (288 hours/year), Senior GIS Analyst, Geographic Information Coordination Program, will contribute as lead analyst on administration of the NC OneMap inventory and in collection of community data holdings. Currently, he is coordinating a statewide GIS data inventory for NC OneMap, and is also heavily involved in the CGIA GIS Coordination Program as lead liaison to the State Government Users Committee and as staff to the Statewide Mapping Advisory Committee. Mr. Giordano, who has worked with CGIA since 1992, is the database manager for the NC Corporate Geographic Database.

Primary Staff (project funded in part)

Julia Harrell (518 hours/year), Applications Programmer, will be the lead applications programmer for development of NC OneMap infrastructure, including the inventory component that will provide data availability information to the proposed project. Ms. Harrell works with various programming tools including ArcIMS, Visual Basic, ArcView Avenue, MapObjects, and HTML. She plays a key role in the Floodplain Mapping Information System and is redesigning the NCMapNet Internet mapping application. Julia worked for the GIS Unit in the NC Department of Transportation before joining CGIA in 2000.

III. Summary of Project Output

III.A Milestones

Acquisition of NC OneMap Data Inventory (version 1) content	Month 2
Hire and train staff	Months 1-3
Determination of layer priority/frequency for time series	Month 4
Cumulative analysis of GICC survey and other surveys complete	Month 4
Disk storage system deployed	Month 6
Digital repository software installation and configuration complete	Month 8
Stage 2 GICC survey information acquisition begun	Month 18
Completion of first full year cycle of state and county vector time series capture	Month 18
Report to Library of Congress on Data Selection and Identification Phase	Month 18
50% of available county orthophoto data acquired	Month 24
Report to Library of Congress on Content Acquisition phase	Month 24
Report to Library of Congress on Partnership Building Phase	Month 30
Report to Library of Congress on Content Retention and Transfer Phase	Month 36
Project Evaluation completed	Month 36

III.B Summary of Deliverables

- Comprehensive archive of county and state digital orthophoto resources
- Archive of selected digital map resources
- Archive of selected state and local vector data resources, including time series data where appropriate
- Archive of selected state and local tabular data resources, including time series data where appropriate (these data may be previously integrated with vector data)
- Documented expertise and best practices in the application of METS (Metadata Encoding and Transfer Standard) to the geospatial data industry for long-term data management
- Documented expertise and best practices with regard to management of geospatial data within emerging open source digital repository software
- Documented expertise and best practices with respect to automated identification and extraction of data from OpenGIS specification-based services
- Cultivation of a greater awareness of time series, long-term access, and digital preservation issues in the data producer community

III.C Evaluation Measures

- Number of counties, municipalities, lead regional organizations, state agencies, and other entities from which data has been acquired
- Total file quantity and data file size acquired
- Percent of data objects for which at least minimal FGDC metadata has been developed
- Percent of data objects for which METS metadata has been developed
- Percent of acquired data objects submitted to the digital repository
- Percent of vector and tabular data resources targeted for time series capture for which continuous time series have been acquired
- Rate of adoption in the state and local geospatial data producer community of preservation strategies and time series development
- Percent of NC OneMap data sources exposing time series data for access